联合作答时间的等级得分模型开发及其应用

刘志城[1]，秦春影[1,2]，罗照盛[1]，喻晓锋[1*]，彭亚风[1]

(1) 江西师范大学心理学院

(2) 南昌师范学院数学与信息科学学院

摘 要 国内外融合作答时间的测量模型研究多以 0-1 计分为基础。然而，在实际测验情境下(如数学测验中的多选题、计算题和应用题等)，常采用等级计分方式。本文基于层次模型框架，在等级得分模型(GRM)上融入作答时间信息，构建联合作答时间的等级得分模型 GRM-RT。参考已有实证研究，针对性地设置研究条件，重点考察在不同被试规模与测验长度条件下模型参数的估计返真性。并进一步将新模型应用于实证数据，一方面展示新模型的使用，另一方面进行不同模型的相对拟合比较。结果表明：在各实验条件下，GRM-RT 模型的参数返真性较好且较为稳定；实证数据分析的结果进一步表明模型的实际应用价值。

关键词： 多级记分，GRM 模型，作答时间，建模

1 引言

随着计算机化测验的广泛应用，教育工作者可以收集到除作答得分数据之外的过程性数据。考生在每个测验题目上的作答时间(response time, RT)是一种过程性数据，它不但能够反映考生的答题速度、思考过程以及对题目的投入程度(Schnipke & Scrams, 2005)，而且能够反映题目的测量特征(Marianti et al., 2014)。例如，有研究在个人拟合度评估(person-fit evaluation)中考虑考生的作答时间和作答得分，结果表明作答时间可以在识别异常考生方面提供非常有用的辅助信息(Fox & Marianti, 2016 Sinharay & Johnson, 2020)。在变化点分析(change point analysis; Page, 1953)中，Cheng 和 Shao(2022)、钟小缘等人(2022)和 Zhu 等人(2023)使用考生的作答时间来识别考生在作答速度上的异常变化。人格量表研究发现，若个体在某一项目上花费时间明显高于其他项目，其对该项目作答就存在不稳定性，即同一被试再次作答可能出现结果不一致现象(Dunn et al., 1972; Holden & Fekken, 1993)。随着信息技术的发展，对测验

结果的分析将不再局限于传统的作答得分，将作答时间考虑进来，为理解考生在测验中的行为提供了动态的、多维度的视角。

联合作答时间和作答得分，van der Linden(2007)提出的层次框架模型是备受关注的模型之一。与传统依赖性建模有所不同，层次框架模型通过将个体特质水平与项目测量特征的影响分为不同层次进行考虑，使得模型能够在更细致的层面上探索作答时间和作答得分之间的相互作用，这不仅提升了模型的灵活性，还为后续研究提供了可靠的基础(Fox et al., 2007; 郭磊 等, 2017)。一方面，国内外研究中将作答时间纳入 IRT 模型的工作，主要基于 0-1 计分的测验展开建模(比如 Molenaar et al., 2015, 2018)。不同于 van der Linden (2007)在层次模型中被试的能力和速度通过协变量建模，在给定考生能力和速度后，考生的作答得分和作答速度无关，而 Molenaar 等人(2015)采用交叉关系函数来量化考生能力和作答时间、速度和作答得分的关系。另一方面，在实际测验场景中，如数学测验中的多选题、计算题和应用题等，通常采用多级得分方式，针对这一需求，已有研究考虑广义部分评分模型(generalized partial credit model, GPCM; Muraki, 1992)联合建模(如郭莹莹,2020; 刘子瑜, 2024)。

等级反应模型(graded response model, GRM; Samejima, 1969)是常见的多级得分模型，相较于 GPCM，在处理有序多级得分数据方面表现出更显著的优势，如常见的 Likert 量表等，在多种模型拟合度指标上，GRM 的表现优于 GPCM(Naumenko, 2014)。借鉴 Molenaar 等人(2015)的思想，Wang 等人(2019)也采用交叉关系函数来处理考生的能力和作答时间，速度和作答得分间的关系，将多维测量模型和作答时间进行联合建模。考生能力和速度对于作答数据的影响按交叉关系函数建模的思路比较适用于健康领域的测量，如 Wang 等人(2019)强调的那样，对于心理和教育测评领域，更宜将考生能力和速度按条件独立关系建模。

随着国家教育数字化战略行动的纵深推进，心理和教育测评领域已经成为教育数字化的重要实施阵地，涉及到等级得分的应用场景很多，比如心理测量领域常见的 Likert 类量表(LaHuis et al., 2011; Beck et al., 2019)，有研究将 GRM 融入到认知诊断测验中(Ma & de la Torre, 2016; Sun et al., 2013)等。计算机化测验的普及以及网络在线的问卷调查平台的兴起，收集被试的作答时间已经成为常规操作，亟须针对不同测验场景，联合多级作答数据和作答时间数据建模，从而为心理和教育测评领域的数据分析提供强有力的工具。

综合来看，现有文献未能充分探讨等级作答得分和作答时间这一结合在实际应用场景中的潜力，尤其是聚焦心理和教育测量领域，将考生能力和速度按条件独立关系建模，联合等级作答得分和作答时间构建的层次 GRM-RT 模型值得深入探索和研究，探索其在多级得分场景中的表现，具有重要的理论意义和实践价值。基于上述考虑，本研究拟基于层次建模框

架，将 GRM 联合作答时间进行建模，开发融入作答时间的 GRM 模型，为多级计分测评数据提供新的理论基础和模型应用场景。

2 GRM-RT 模型的开发

为了方便介绍，在这里首先给出本文所涉及到的符号及其含义，如下表所示。

表 1 本文所涉及到的符号及其含义

符号	含义	符号	含义
$\boldsymbol{T}$	考生作答时间矩阵	$\boldsymbol{Y}$	考生作答得分矩阵
i	考生的下标	j	题目的下标
$\boldsymbol{\theta}$	考生能力参数	$\boldsymbol{\tau}$	考生速度参数
$\boldsymbol{a}$	项目区分度参数	$\boldsymbol{b}$	项目难度参数
$\boldsymbol{\alpha}$	项目时间区分度参数	$\boldsymbol{\beta}$	项目时间强度参数
$\boldsymbol{\varepsilon}$	对数正态时间模型的残差项	$\boldsymbol{t}$	项目的难度等级
ρ	考生能力和速度参数的相关性	$\boldsymbol{\mu}$	均值向量
$\boldsymbol{\sigma}$	标准差	$\boldsymbol{\Sigma}$	协方差矩阵

下面将分别介绍本文中需要用到的对数正态作答时间模型(van der Linden, 2006)和等级得分模型(Samejima, 1969)，之后将探索将它们融合以构建联合作答时间的层次等级得分模型(GRM-RT)。

2.1 对数正态作答时间模型(lognomal-response time, LN-RT)

van der Linden(2006)提出的 LN-RT 认为，作答时间服从对数正态分布，并假设在个体潜在能力值恒定的情况下，作答时间仅受到考生作答速度和项目时间参数的影响，以固定的速度对每个项目作答。模型可以表达为如下的公式 1:

$$f\left(T_{ij};\tau_i,\alpha_j,\beta_j\right)=\frac{\alpha_j}{T_{ij}\sqrt{2\pi}}\exp\{-\frac{1}{2}\left[\alpha_j\left(\ln T_{ij}-\left(\beta_j-\tau_i\right)\right)\right]^2\}, \quad (1)$$

公式 1 还可以表达为下面的形式：

$$\log(T_{ij})=\beta_i-\tau_j+\varepsilon_{ij}, \quad (2)$$

其中，T_{ij}表示被试i在第j题上的作答时间，τ_i表示被试i的加工速度参数，α_j表示第j题的时间区分度参数，α_j值越大区分考生作答速度能力越强，β_j表示第j题的时间强度参数，β_j越大表示被试在该项目要花费更长的时间，ε_{ij}为正态分布的误差项。

由于 LN-RT 模型对测验数据有较好的拟合表现，很多研究者基于这个模型展开了深入的探索和研究，比如：作答时间的测量模型开发(Liu & Wang, 2024; 田亚淑 等, 2023; Ulitzsch et al., 2022; Wang & Xu, 2015; 詹沛达 等, 2020),使用作答时间对考生的异常行为进行检测(Han & Kang, 2021; Shao & Cheng, 2022; Zhu et al., 2023),在自适应测验中使用作答时间选题(van Der Linden, 2008; Choe et al., 2018)等。

2.2 等级反应模型(GRM)

Samejima(1969)提出了等级反应模型，自提出该模型，它在测量领域受到广泛关注和应用(Beck et al., 2019; Ma & de la Torre, 2016; Sun et al., 2013; Yu & Cheng, 2019)。在 GRM 下，将题目分为多个得分等级，等级难度是随着等级逐级递增的。记 $P_{aj,t}$表示考生a在第j题得t分的概率，$P^*_{aj,t}$表示考生a在第j题上得t分及t分以上的概率，则

$$P_{aj,t} = P^*_{aj,t} - P^*_{aj,t+1}, \tag{3}$$

其中:

$$P^*_{aj,t} = \frac{1}{1 + \exp(-Da_j(\theta_i - b_{jt}))}, \tag{4}$$

$$P^*_{aj,0} = 1, P^*_{aj,m_i+1} = 0, \tag{5}$$

假设题目j的满分为f_j分, 那么可以分为$f_j\ +1$个得分等级，即$0,1,2,\dots,f_j$。$P^*_{aj,t}$ 表示能力为θ_i的考生在第j道题目上得分$\geq t$ 分的概率，$P_{aj,t}$表示被试在第j道题目上恰得t分的概率。b_{jt}表示第j道题上得t分的难度，其难度等级单调递增,$b_{j1} < b_{j2} <\cdots b_{jt}$。

鉴于 LN-RT 和 GRM 模型分别在作答时间和作答得分数据上的良好表现和广泛应用，本研究考虑将它们融合构建可以处理作答时间和多级得分的数据，下面介绍具体的模型构建思路和过程。

2.3 GRM-RT 模型

不同于 Molenaar 等人(2015)和 Wang 等人(2019)将考生的能力和速度考虑交叉效应，本研究面向常见的心理和教育测量情景，故仍然沿用 van der Linden (2007)中的层次结构，即考生特质(能力和速度)对于作答(作答得分和作答时间)按条件独立建模，使用 GRM 和 LN-RT 来联合构建层次 GRM-RT 模型。

在层次 GRM-RT 模型中，将考生的作答得分Y和作答时间在第一层建模，而在第二层中，将有关作答得分和作答时间的项目参数和被试参数分别进行建模。通过层次框架，GRM-

RT 模型在关注作答时间和作答反应的条件独立性的同时，也捕捉了它们整体上的关联性。而项目参数和被试参数之间的相关性则放在最高层，通过均值向量和协方差矩阵进行建模。

根据层次框架模型，GRM-RT 的被试参数和项目参数服从多元正态分布假设，给定均值向量和协方差矩阵，则它们的分布可以表示为:

$$(\boldsymbol{\theta},\boldsymbol{\tau})^T \sim N_2(\boldsymbol{\mu}_p,\boldsymbol{\Sigma}_p), \tag{6}$$

$$(loga_j, b_j, \alpha_j, \beta_j)^T \sim N_4(\boldsymbol{\mu}_I,\boldsymbol{\Sigma}_I), \tag{7}$$

其中$\boldsymbol{\mu}_p$和$\boldsymbol{\Sigma}_p$分别代表被试参数的均值向量和协方差矩阵，μ_I和Σ_I分别代表项目参数的均值向量和协方差矩阵。

在模型的参数识别性方面，被试能力参数$\boldsymbol{\theta}$服从标准正态分布假设，这种标准化确保了能力参数的唯一性，并防止了由于尺度差异导致的不可识别性问题(van der Linden, 2007)。在速度参数τ的生成过程中，设置$\sigma_\tau = \sigma_{speed} \times \sqrt{1-\rho^2}$，其中$\rho$表示了速度参数和能力参数之间的相关关系，$\sigma_\tau$对速度的波动进行了有效约束，使得速度的变化独立于能力参数的方差，进一步确保了模型的稳定性和识别性(Patton, 2015)。其次，给定作答时间模型的参数后，$\log(\boldsymbol{T}_i)$满足局部独立性假设，给定作答模型的参数后，被试作答$\boldsymbol{Y}_i$也满足独立性假设(van der Linden, 2007)。

结合公式 6 和公式 7，GRM-RT 的模块框架可以由下面的图 1 表示:

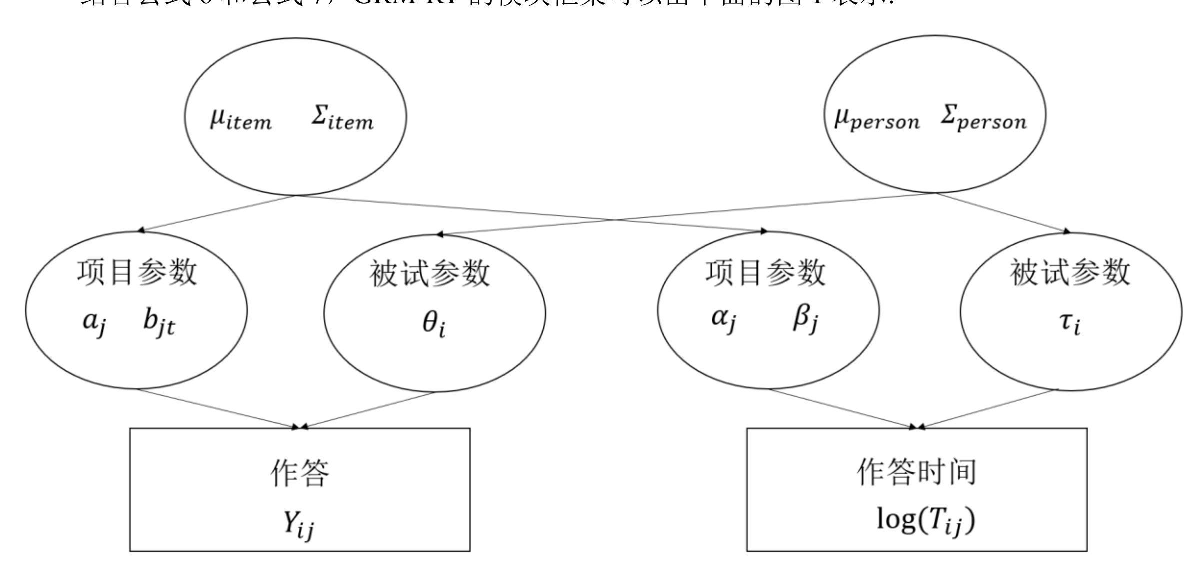


图 1 层次 GRM-RT 模型框架图

可以看出，新构建的层次 GRM-RT 沿用了 van der Linden(2006, 2007)相似的模型框架，并且也遵循了考生作答得分和作答时间条件独立性假设，即给定考生能力时，考生在各题目上的作答时间只与考生的速度有关系；给定考生速度时，考生在各题目上的作答得分只与考生的能力有关系，这样的设置更符合心理和教育测量场景(Wang et al., 2019)，同

时也会带来一些额外的好处，比如(1)层次模型框架允许独立地分析考生作答得分和作答时间，也可以将它们进行联合分析；(2)可以实现混合计分(测验同时包括二级得分题目和多级得分题目)时的数据分析，因为二级计分可以看作是 GRM 的特例(陈青, 2008)。

3 GRM-RT 模型的参数估计

为了考察新构建的 GRM-RT 模型在参数返真性上的表现，参考已有研究中的有关设置(Wang et al., 2013; Patton, 2015)，在较理想的条件下和常规条件下展开如下的模拟实验。

3.1 模型的参数设置

在被试参数方面，能力参数θ服从标准正态分布即$\theta \sim N(0,1)$，速度参数τ服从正态分布，均值和标准差的设置如下：参考 Wang 等人(2013)对实证数据分析的结果，能力参数θ和速度参数τ之间的相关系统ρ设置为0.5，这种相关设计符合实际的情况，即能力越强的考生作答速度也相应更快。其次将$\sigma_{speed} = 0.25$视为速度的参数基础的标准差，速度参数τ的均值设定为$\mu_\tau = \rho \times \theta \times \sigma_{speed}$，速度参数$\tau$的标准差设定为$\sigma_\tau = \sigma_{speed} \times \sqrt{1-\rho^2}$。这种设置保证了能力参数$\theta$和速度参数$\tau$的相关性，同时 σ_{speed}让参数在合理范围内波动(Patton, 2015)。

在项目参数方面，区分度参数的设置为$a \sim logN(0,0.5)$，难度参数将最容易的难度等级设置为 $b_{j1} \sim N(-1.5,0.5)$，为了保证难度递增设置了Δb参数，Δb设置为$\Delta b \sim N(1,0.2)$，所以 $b_{j2} = b_{j1} + \Delta b$，后续的难度设置依次递增为 $b_{jt} = b_{j1} + t \cdot \Delta b$ (Manapat & Edwards, 2022)。时间区分度参数的设置为$\alpha \sim U(1.5,2.5)$。参考 Patton (2015)基于实证数据分析得到的研究设计：即基于回归模型生成时间强度参数β，设置期望均值为$\mu_\beta = 4$，期望的标准差为$\sigma_\beta = 0.3$，β与a的相关性为 $\rho_{a\beta} = 0.3$，β与b的相关性为$\rho_{b\beta} = 0.5$，同时计算回归系数Z_a和Z_b、截距项C、以及误差项的方差σ_ε。具体计算公式为:

$$Z_a = \frac{\sigma_\beta}{\sigma_a} \times \rho_{a\beta}, \qquad Z_b = \frac{\sigma_\beta}{\sigma_b} \times \rho_{b\beta}, \tag{8}$$

$$C = \mu_\beta - (Z_a \times \mu_a + Z_b \times \mu_b), \tag{9}$$

$$\sigma_\varepsilon = {\sigma_\beta}^2 - ((Z_a \cdot \sigma_a)^2 + (Z_b \cdot \sigma_b)^2 + 2 \cdot Z_a \cdot Z_b \cdot \sigma_{ab}), \tag{10}$$

其中μ_a和μ_b分别为所有项目参数a和中等难度b的总体均值，σ_a和σ_b分别为所有项目参数a和中等难度b的总体标准差，σ_{ab}代表所有项目参数a和中等难度b的协方差。所以β的生成服从条件正态分布为$\beta \sim N(C + Z_a \cdot a + Z_b \cdot b, \varepsilon), \varepsilon \sim N(0, \sigma_\varepsilon)$。这种设计使生成的$\beta$参数既满

足了多个参数之间的相关性，通过设置标准差让参数可以在合理范围内波动，能够很好满足数据生成的需求。

参数估计采用 Hamiltonian Monte Carlo(HMC; Betancourt, 2017)方法进行。HMC 方法是一种改进的马尔科夫链蒙特卡洛(MCMC)方法,与传统的随机游走 MCMC 方法(Metropolis-Hastings)相比，HMC 通过利用目标分布的梯度，能够以更少的迭代次数探索到目标分布的高概率区域，从而显著降低了模拟的计算成本(Betancourt, 2017)。具体的 rstan 模型代码见附录二。

3.2 研究设计

本研究的目的是为了评估层次 GRM-RT 模型的参数返真性和稳健性，采用 2×2×3 的三因素实验设计，自变量分别是被试能力分布、被试人数、测验长度。在实际的应用中，特别是样本较小时，考生的能力分布往往不符合标准正态分布，因此考虑被试能力分布有正态分布和偏态分布两个水平，被试人数有 200、500 和 1000 人三个水平，测验长度有 10、20、40 题三个水平。所有题目为 5 个等级的多级计分，参照 Kieftenbeld 和 Natesan(2012)的偏态分布生成方法，使用 Fleishman 变换法将能力参数θ转换为偏态分布下的θ'。设置$k_1 = -0.282, k_2 = 1.037, k_3 = 0.282, k_4 = -0.042$，通过等式$\theta' = k_1 + k_2\theta + k_3\theta^2 + k_4\theta^3$来进行转换。具体的分布形式如图 2 所示，可以看出中等水平附近的学生数量最多，其次占比较多是能力较高水平的学生，而能力特别低的考生最少，这种能力分布与很多实证数据中得到的结果相近(Molenaar et al., 2012; Trafimow et al., 2019; Veldkamp et al., 2017)。

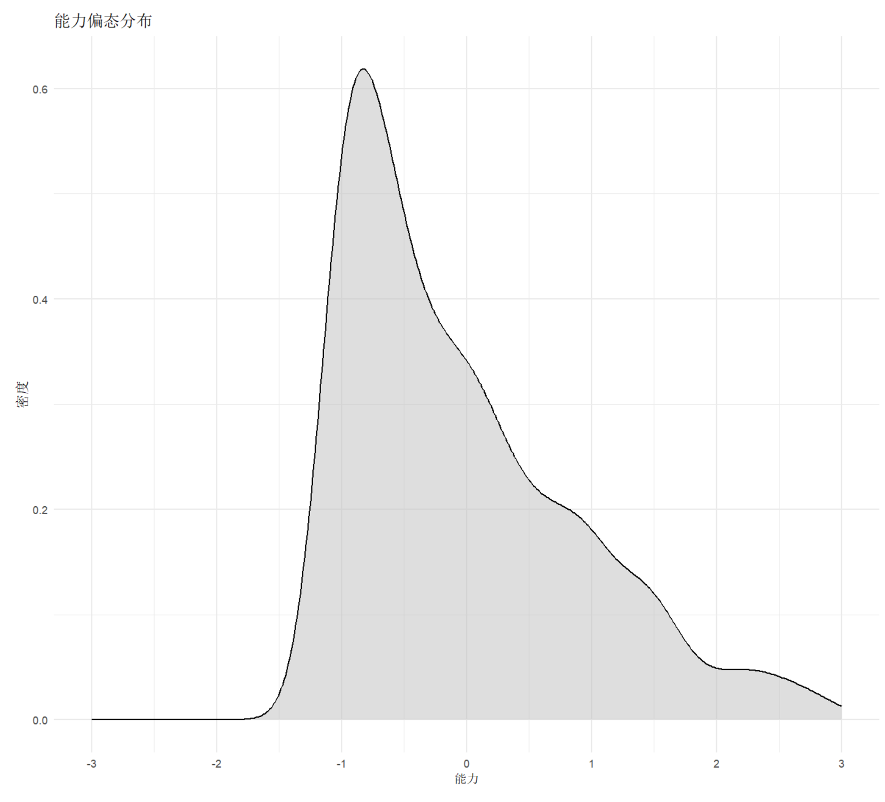


图 2 偏态分布图像

3. 3 参数估计

整个过程分为四步，第一步首先使用 3.1 模型参数设置的方法生成出被试参数和项目参数，再将得到的参数真值代入到 GRM 和对数正态作答时间模型中，得到被试的作答矩阵和作答时间矩阵。

第二步将之前生成的作答数据输入到 R 软件的自编代码进行参数估计，在这一阶段，HMC 方法进行参数估计主要依赖于先验分布和似然函数，其中各个参数的先验分布设置见表 1。在模型拟合过程中，HMC 方法对后验分布进行采样。每一次的迭代次数为 6000 次，燃烧期为 3000 次，2 条链同时进行。

最后将参数的估计值和参数的真值进行比较，使用均方根误差(Root Mean Square Error, RMSE)和平均离差(Bias)两个指标对模型的估计精度进行评估。为了减少随机误差的影响，重复上述步骤 30 次。

表 2 参数的先验分布

参数	先验分布
考生能力参数θ	$N(0,1)$
考生速度参数τ	$N(\mu_\tau, \sigma_\tau)$
区分度参数a	$LogN(0,0.5)$
难度参数b_t	$N(-1.5,0.5) + t \cdot \Delta b$

时间区分度参数α	$N(2,0.167)$
时间强度参数β	$N(4,0.3)$

3.4 评价指标

为了衡量估计的精度，评价指标使用了均方根误差和平均离差，其中均方根误差(RMSE)用于衡量参数估计值与真实值之间的平均偏差，其数值越小意味着参数估计的精度越高。

$$RMSE(\hat{v}) = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(\hat{v}_n - v_n)^2}\ , \tag{11}$$

平均离差(Bias)用于描述参数估计值相对于真实值的偏离程度。偏离程度越小，意味着估计的准确性越高。

$$Bias(\hat{v}) = (\sum_{n=1}^{N}\hat{v}_n - v_n)/N, \tag{12}$$

其中$\hat{v}_n$表示参数的估计值，v_n表示参数真值。

3.5 结果与分析

3.5.1 参数收敛性判断

在参数估计的收敛性方面，以研究中最小样本量及测验长度的条件为例，观测各个参数的收敛情况。具体来说，使用正态分布下被试人数 200 人以及测验长度为 10 题的条件。首先通过马尔科夫链监测轨迹图(trace plot)来判断参数的收敛情况(Hung & Wang, 2012)。在上文的设计中，参数迭代次数为 6000 次，前 3000 次作为燃烧期，关注后面 3000 次迭代的轨迹。通过观察参数的轨迹图(即参数值随迭代次数的变化曲线)，我们可以通过判断轨迹是否围绕某个水平上下波动，这表明链已经达到了平稳状态。如果链还在不断偏移某个方向，则表明还没有达到收敛，需要继续采样。图 3 是在最小样本量及测验长度的条件下各个参数的轨迹图，从轨迹图中可以看出各个参数均围绕某个水平波动，说明参数已经收敛。$\hat{R}$(Rhat)一种用于评估马尔科夫链蒙特卡洛算法收敛性的指标，由 Gelman 和 Rubin 提出，Brooks 和 Gelman (1998) 中进一步改进。它通常被称为潜在尺度缩减因子(Potential Scale Reduction Factor, PSRF)。通过 2 条链计算的各个参数的$\hat{R}$值都小于 1.1，且接近 1，进一步说明了参数估计算法已成功收敛。

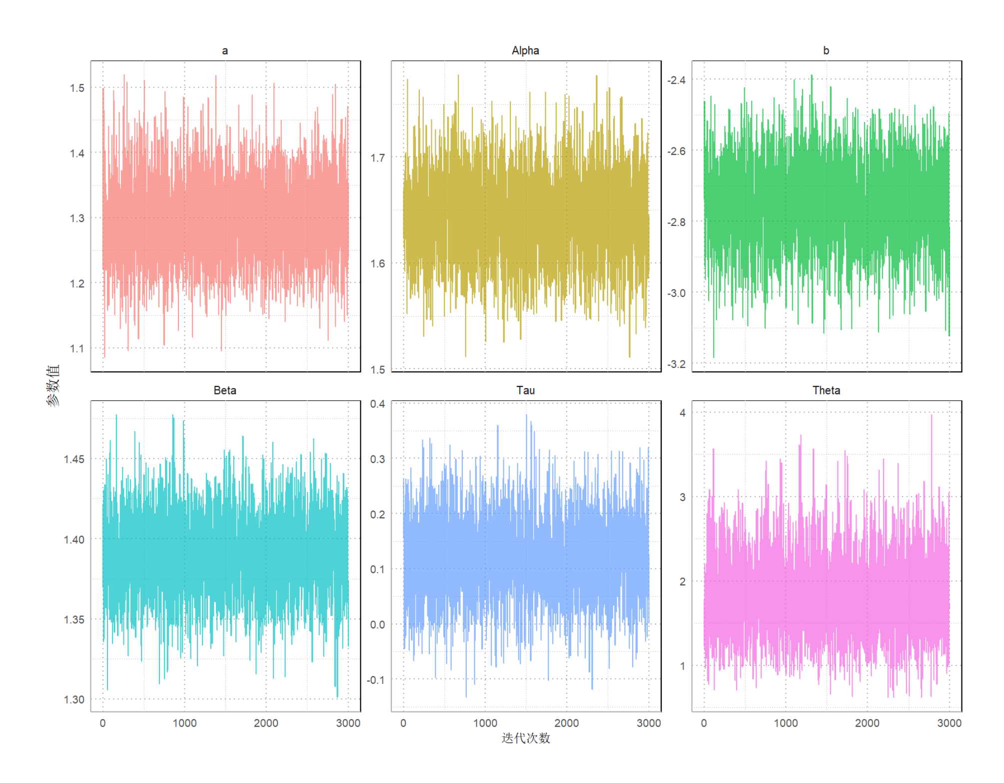


图 3 参数轨迹图

3.5.2 参数估计结果分析

在对 GRM-RT 模型参数估计的收敛性验证后，接下来使用均方根误差和平均离差这两个指标对不同能力分布、不同被试数量及不同测验长度的条件下对测验的准确性进行评估。其中表 2 到表 5 分别是不同条件下使用 GRM-RT 模型估计项目参数和被试参数的返真性情况。

表 3 被试参数正态分布下 GRM-RT 模型的被试参数返真性检验

实验条件	θ		τ	
	RMSE	Bias	RMSE	Bias
$N=200, I=10$	0.242	-0.011	0.062	-0.012
$N=200, I=20$	0.184	0.017	0.055	0.007
$N=200, I=40$	0.149	-0.035	0.036	-0.013
$N=500, I=10$	0.258	0.009	0.065	-0.004
$N=500, I=20$	0.189	-0.008	0.056	0.007
$N=500, I=40$	0.157	0.038	0.039	-0.014
$N=1000, I=10$	0.268	-0.004	0.068	0.017
$N=1000, I=20$	0.205	-0.007	0.058	-0.008
$N=1000, I=40$	0.147	-0.005	0.037	-0.006

注: *N* 为被试人数，*I* 为测验长度。

从表 3 的结果可以看出，一方面，在固定测验长度时，三种考生人数(200、500 和 1000)下的被试参数(θ和τ)的精度相近；在固定考生人数时，被试参数的精度会随着测验长度的增

加而变好。结果表明，相对于被试人数，测验长度对于被试参数的估计精度影响更明显。另一方面，相对于能力参数，速度参数的估计精度更好，这主要是因为作答时间是连续数据，可以包含更丰富的信息，与以往的研究结果一致(郭莹莹, 2020)。

表 4 被试参数正态分布下 GRM-RT 模型的项目参数返真性检验

实验条件	a		b		α		β	
	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias
$N = 200, I = 10$	0.108	-0.033	0.140	0.028	0.088	-0.024	0.027	-0.018
$N = 200, I = 20$	0.119	-0.016	0.173	0.036	0.087	0.039	0.036	0.011
$N = 200, I = 40$	0.123	0.031	0.146	0.068	0.100	-0.003	0.039	-0.024
$N = 500, I = 10$	0.084	-0.028	0.118	-0.009	0.051	-0.006	0.022	-0.006
$N = 500, I = 20$	0.062	-0.012	0.112	0.019	0.057	0.022	0.031	0.017
$N = 500, I = 40$	0.089	-0.005	0.109	-0.011	0.063	0.007	0.039	-0.030
$N = 1000, I = 10$	0.064	0.011	0.064	0.007	0.041	0.006	0.043	0.039
$N = 1000, I = 20$	0.087	0.012	0.085	-0.005	0.039	0.019	0.026	-0.015
$N = 1000, I = 40$	0.060	0.016	0.081	0.022	0.043	0.013	0.022	-0.005

注: N 为被试人数，I 为测验长度。

表 4 列出了被试参数正态分布下 GRM-RT 模型的项目参数估计精度，总的来说，项目参数的精度比对应的被试参数精度更好。在固定测验长度或者固定被试人数时，项目参数的估计精度较好，但是并没有体现出明显的规律，这表明，从样本量来说，500 人对于 GRM-RT 模型已经可以较好的恢复项目参数。

表 5 被试参数偏态分布下 GRM 模型的被试参数返真性检验

实验条件	θ		τ	
	RMSE	Bias	RMSE	Bias
$N = 200, I = 10$	0.268	0.015	0.068	0.010
$N = 200, I = 20$	0.197	-0.010	0.061	-0.018
$N = 200, I = 40$	0.144	0.017	0.033	-0.003
$N = 500, I = 10$	0.278	-0.005	0.065	-0.011
$N = 500, I = 20$	0.189	0.042	0.056	0.005
$N = 500, I = 40$	0.141	0.019	0.037	-0.005
$N = 1000, I = 10$	0.270	-0.006	0.068	-0.013
$N = 1000, I = 20$	0.210	0.016	0.057	0.008
$N = 1000, I = 40$	0.148	0.011	0.039	-0.012

注: N 为被试人数，I 为测验长度。

表 6 被试参数偏态分布下 GRM 模型的项目参数返真性检验

实验条件	a		b		α		β	
	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias
$N = 200, I = 10$	0.145	-0.067	0.161	0.029	0.083	-0.038	0.022	0.010

$N = 200, I = 20$	0.123	-0.021	0.146	0.004	0.093	0.044	0.048	0.032
$N = 200, I = 40$	0.138	0.031	0.164	0.038	0.093	-0.007	0.033	-0.003
$N = 500, I = 10$	0.064	-0.006	0.109	0.010	0.044	0.004	0.030	-0.022
$N = 500, I = 20$	0.102	-0.028	0.108	0.056	0.054	0.019	0.030	0.016
$N = 500, I = 40$	0.081	-0.005	0.118	0.039	0.068	0.003	0.029	-0.013
$N = 1000, I = 10$	0.088	-0.038	0.082	0.011	0.059	0.011	0.027	-0.022
$N = 1000, I = 20$	0.074	0.006	0.084	0.024	0.042	0.020	0.026	0.015
$N = 1000, I = 40$	0.064	-0.022	0.093	0.036	0.043	0.011	0.028	-0.018

注: N 为被试人数，I 为测验长度。

相同被试人数和测验长度条件下，表 5 和表 6 中对应参数的估计精度与表 3 和表 4 中的结果相近，被试特质参数的偏态分布并不会影响被试参数和项目参数的估计精度，表明 GRM-RT 模型的稳健性较好。

综合表 3 和表 5 的结果，被试参数的估计结果良好，均达到较理想的水平，即使在样本量较小(200)和短测验条件(测验长度为 10)下也能保持较高的精度。这表明，即使在资源有限的情况下，所采用的估计方法依旧具备可靠性和稳定性。随着测验长度的增加，估计精度逐步提升，而样本量的增加对精度提升影响较小。这与 Kieftenbeld 和 Natesan (2012) 的研究结果一致：被试人数从 150 增加到 300 时，参数估计的准确性显著提高；进一步增加到 500 时，效果提升较小；从 500 增加到 1000 时，改善更加有限。同时，样本量与测验长度之间的交互作用较小，对被试参数的影响几乎可以忽略，表明增加测验长度或被试人数中的任何一个因素对另一因素的依赖性不强。此外，即使在偏态分布下，考生参数的估计精度依旧良好，说明模型具有很好的稳健性和适应性，能够在不同的分布条件下保持稳定的估计性能。这一点对于实际应用具有重要意义，尤其是在无法保证数据正态分布的情况下，模型依旧能够提供可信的估计结果。

同样从表 4 和表 6 的结果可以得出，项目参数的估计精度表现良好，在正态分布各个条件下参数估计的变化范围非常小，主要受到随机误差的影响。在偏态分布下，参数估计的精度依旧保持高水平，说明模型在不同分布条件下依然稳健可靠。一方面，本研究中采用的中等样本量(分别是 500 和 1000)，对于增加考生人数并没有对项目参数精度带来较多的贡献，其主要的原因是样本量的增加效益呈现边际递减，从 500 人增加到 1000 人时提升幅度很小；另一方面，三种测验长度(10, 20 和 40)下，最短测验长度下的精度相对较低，而当测验长度达到 20 或以上时，测验长度对项目参数估计精度的影响较小，效应大小一般小于 1.65%(Kieftenbeld & Natesan, 2012)。

总的来说，GRM-RT 模型参数估计的误差较小，可以很好的结合作答时间信息对数据进行测量。模型在不同条件下表现稳健，无论是正态分布还是偏态分布，模型都能提供可靠的参数估计。为了进一步考察新模型在实际测验数据中的表现，我们开展了如下的实证研究。

4 实证研究

实证研究在真实数据下评估了 GRM-RT 模型在实际场景中的使用。这里使用的数据集和问卷与刘子瑜(2024)实证研究中所使用的一致，因此可以对 RTs-mGPCM 模型进行直接比较。测验项目属于教育领域，测量的主题是数学素养测试。

表 7 人口统计学变量

	类别	频次	百分比(%)
性别	男	404	47.9
	女	439	52.1
年级	初一	416	49.3
	初二	208	24.6
	初三	219	25.9

被试的总人数为 843 人，均为在校初中生。具体人口统计学的信息见表 7。测试的目的是调查初中生的数学素养，共包含二十个项目，采用李克特四点计分的形式考察学生对于正确回答该项目问题的确信程度，即 1=非常没把握答对，2=不太有把握答对，3=有把握答对，4=非常有把握答对，项目的具体内容见附录一。

4. 1 研究设计

实证数据同样采用 HMC 方法进行参数估计，马尔科夫链迭代次数为 6000 次，燃烧期设置为 3000 次，2 条链同时进行，同样使用自编的 R 语言代码完成分析。

4. 2 评价指标

由于真实数据没有被试参数和项目参数的真值，无法计算各参数的 RMSE 和 Bias，因此这里与刘子瑜(2024)一致，使用广义适用信息准则(Widely Applicable Information Criterion, WAIC; Watanabe & Opper, 2010)和留一交叉验证(Leave-One-Out, LOO; Vehtari, Gelman, & Gabry, 2017)作为评价指标。WAIC 是一种贝叶斯模型选择指标，被认为是类似 AIC 的一种广义形式，适用于更复杂的模型，衡量了模型的对数似然的期望，以及对模型预测能力的惩罚。LOO 是一种交叉验证的方法，用于评估模型的预测能力。在贝叶斯统计中，LOO 可以通过对每一个数据点进行“留一法”交叉验证来计算每个样本在不包含该样本时的对数似然。两个指标的具体的公式如下:

$$WAIC_{ELPD} = \sum_{i=1}^{N} log\left(\frac{1}{s}\sum_{s=1}^{s} p(y_i|\gamma_s)\right) - \sum_{i=1}^{N} V^s(logp(y_i|\gamma_s)), \tag{13}$$

$$LOO_{ELPD} = \sum_{i=1}^{N} \log(\frac{1}{s}\sum_{s=1}^{s} \frac{p(y_i|\gamma_s)}{1-p_i(\lambda_s)}), \tag{14}$$

其中N为样本数量，y_i表示第i个观测值，s表示后验样本的数量，γ_s表示第s个后验样本，V^s表示对数似然的方差，$p_i(\lambda_s)$表示每个样本点对后验分布的贡献。通过 WAIC 和 LOO 指标来判断 GRM-RT 模型相较于 GPCM 版的模型(刘子瑜, 2024)拟合的结果是否具有优势，这两个指标都是值越小表明模型的相对拟合越好。

表 8 模型拟合指标

组别	指标	GRM-RT	RTs-mGPCM
性别	WAIC	23457.7	47399.6
	LOO	23466.6	46518.5
年级	WAIC	23142.3	46290.8
	LOO	23152.9	45496.9

对于这批数据，刘子瑜(2024)对被试进行了分组检验，将被试的性别和所在年级进行了统计。为了与刘子瑜(2024)的 RTs-mGPCM 模型进行比较，我们根据数据标签将被试分为年龄组和年级组。随后，使用 GRM-RT 模型进行估计，并采用 WAIC 和 LOO 指标评估模型。两个模型的拟合指标结果见表 8 所示。可以看出，GRM-RT 模型的拟合指数优于 RTs-mGPCM，GRM-RT 的 WAIC 和 LOO 值均为 RTs-mGPCM 的一半左右，这两个指标都表明，本研究中提出的模型对于这批实证数据有更好的拟合。此外，不考虑分组的情况下，WAIC 和 LOO 的值分别为 23314.4 和 23321.8。模型对于集体分析还是考虑组别差异的情况下均保持高拟合，说明模型可以很好的适用于教育测量领域。

表 9 项目参数估计

	a	b_1	b_2	b_3	α	β
项目 1	1.2301	-2.6217	-1.2112	0.4647	1.5040	2.5170
	(0.001)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)
项目 2	1.1913	-2.5742	-1.1829	0.6137	1.5052	1.6639
	(0.001)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)
项目 3	1.3312	-2.7470	-1.3747	0.2598	1.5966	1.5181
	(0.001)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)
项目 4	1.2610	-2.8019	-1.4199	0.4297	1.6754	1.2314
	(0.001)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)
项目 5	1.2950	-2.7486	-1.3285	0.1174	1.6422	1.3912
	(0.001)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)

项目 6	1.2179	-2.6529	-1.2510	0.2875	1.7692	1.2599
	(0.001)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)
项目 7	1.5376	-2.6595	-1.3000	0.1097	1.8299	1.1353
	(0.001)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)
项目 8	1.7077	-2.6061	-1.6029	-0.2731	1.9070	1.0237
	(0.001)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)
项目 9	2.0487	-2.6482	-1.5074	-0.2471	1.7255	0.9530
	(0.001)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)
项目 10	1.3329	-2.7278	-1.3902	0.2569	1.8106	1.0028
	(0.001)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)
项目 11	1.0281	-2.4667	-1.0055	0.4792	1.8134	1.0999
	(0.001)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)
项目 12	1.3065	-2.8213	-1.4607	0.0143	1.5204	1.2665
	(0.001)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)
项目 13	1.5348	-2.5864	-1.3412	0.1834	1.7524	0.9187
	(0.001)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)
项目 14	1.1894	-3.2136	-2.0889	-0.5155	1.6739	1.0604
	(0.001)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)
项目 15	1.1035	-2.4015	-0.7308	0.8300	1.5988	1.3056
	(0.001)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)
项目 16	1.1133	-2.2700	-0.8268	0.5363	1.7594	1.0776
	(0.001)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)
项目 17	1.6561	-2.8502	-1.6453	-0.2088	1.8341	0.9099
	(0.001)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)
项目 18	1.6241	-2.4212	-1.1229	0.1875	1.7309	1.0786
	(0.001)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)
项目 19	1.6877	-2.4638	-1.2699	-0.0111	1.7599	0.9096
	(0.001)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)
项目 20	1.7216	-2.5446	-1.3627	-0.0584	1.6225	1.0611
	(0.001)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)

注: a为项目区分度参数，b为项目难度参数，α为时间区分度参数，β为时间强度参数。括号内参数的标准误差(Standard Error, SE)。

表 10 考生特质参数统计量

统计量	θ	τ
μ	-0.032	0.036
σ	1.056	0.289

注: θ是考生的能力参数，τ是考生的速度参数。

同样地，本研究运用 GRM-RT 模型针对问卷数据展开参数估计工作，具体的参数估计结果呈现在表 9 之中。考生特质参数的统计量以及特质分布的直方图见表 10 及图 4，考生

能力和速度参数之间存中低程度的相关，相关系数为 0.226。在实际的应用中，考生与项目的实际参数分布可能与模型预先设定的分布存有差异，各个参数估计较小的标准误表明此模型依旧可以维持较高的估计精度，说明模型具备良好的稳健性。考生能力和速度参数的均值和标准差与模型先验设置的均值和标准差十分接近，说明模型对学生能力总体水平和离散程度的初始判断是合理的。此外，直方图显示的考生能力的分布与图 2 偏态分布描述的信息接近，都是中等水平附近考生最多，其次是较高水平的考生，而较低水平的考生特别少，说明此类偏态分布可以反映中学学生的能力分布情况。在 GRM-RT 模型所设定的分布框架下，从各题目所对应的项目难度以及时间强度数值来看，说明这些项目属于难度较低且时间压力较小的类型。这一设定与问卷实测模式契合，即无需考生在特定时间限制内完成答题，同时也不要求考生计算题目的最终答案。其中，项目 1 的时间强度参数数值最大，而其难度参数显示该项目并非难题，不过大部分考生在项目 1 上耗费了最多时间，这或许是由于热身效应(Shao, 2016)的存在，考生需要先适应正常考试流程，所以才会花费更多时间进行作答。项目 9 的区分度处于最高水平，与刘子瑜(2024)的分析结论相一致，原因在于项目 9 涵盖了初中阶段的大部分知识点，而低年级学生尚未学习这些内容，表现出较高的区分度。

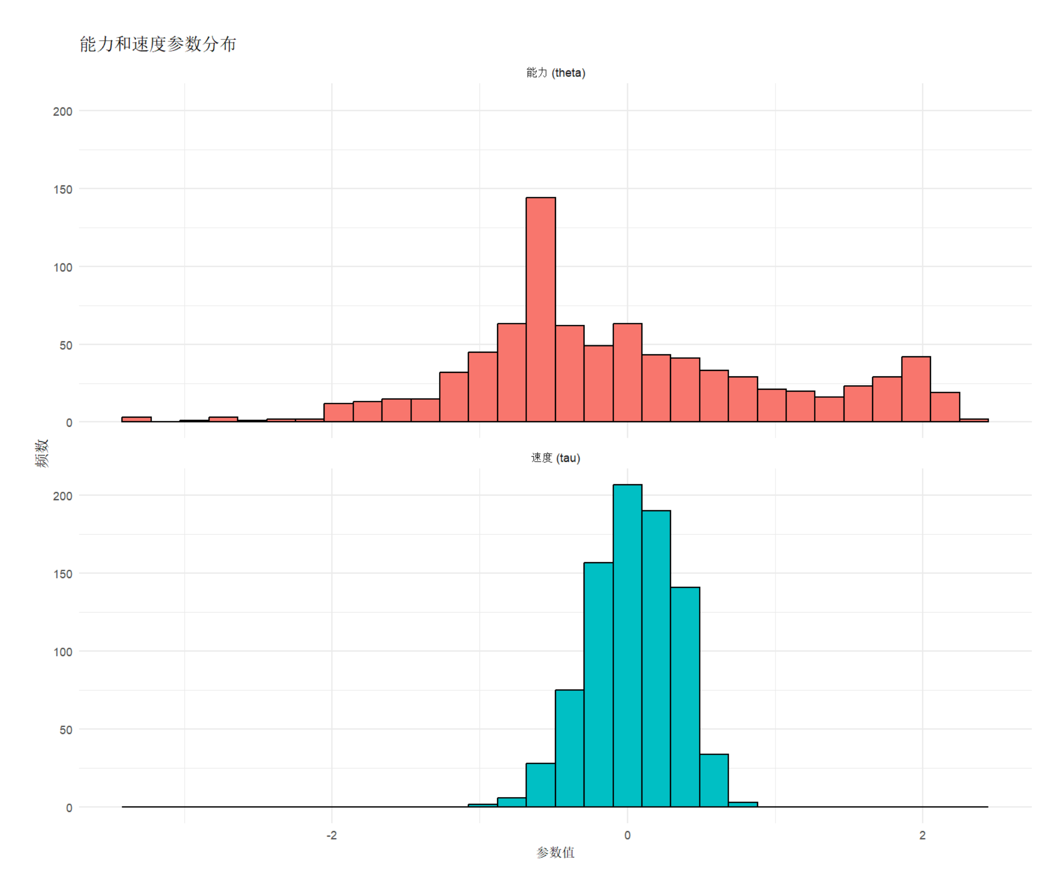


图 4 考生能力和速度参数直方图

5 总结与展望

随着计算机化测验的普及以及网络在线的问卷调查平台的兴起，收集考生的过程作答数据已经不在困难，可以很容易获取在传统纸笔测验上无法收集的过程性数据，如作答时间、鼠标点击次数(De Boeck & Scalise, 2019; Qiao et al., 2023)、眼动轨迹(Man & Harring, 2019; Man & Harring, 2021)等，这些数据可以帮助研究者更好的理解考生思考的过程以及作答行为的转变。国内外的众多研究也将作答时间这类过程性数据纳入 IRT 模型中进行建模，但多数基于 0-1 计分的数据展开建模。然而，我国的实际的测验场景相对更复杂，涉及的题目如量表问卷、多项选择题和主观题等，通常采用多级计分方式。针对这一需求，国内针对作答时间的多级计分模型虽然有一些研究，但是集中于 GPCM 模型，而 GRM 模型在处理有序多级评分数据方面有其特点和优势(Naumenko, 2014)。因此，本研究开发了联合作答时间的 GRM 模型，以期进一步提升测验数据分析的精度和适用性。需要注意的是，不同于前人研究中的考生能力和速度对于作答时间和作答得分按交叉关系建模(Molenaar et al., 2015; Wang et al., 2019)，本研究开发的 GRM-RT 模型考虑的是作答得分和作答时间数据的条件独立，GRM-RT 更适用于常规的心理和教育测量场景。

本研究一方面通过针对性设计的模拟研究对参数的返真性进行了检验，结果表明在不同样本量和测验长度下，模型的各参数均表现出较高的精度。模型在偏态分布的情况下，估计精度也能保持较高的水平，说明模型具有良好的稳健性。在实证研究中分析了教育学领域的数学素养量表，结果表明在模型拟合度方面优于 GPCM 版的模型(刘子瑜, 2024)，进一步说明了模型的测验性能优异且有教育领域的应用前景。模型的构建方式使得 GRM-RT 不限于教育统计和心理测量领域。其它心理和教育学领域，如咨询、社会和发展教育方向等，仍较多使用量表问卷来收集实验数据。而多级作答数据和作答时间数据的联合分析，使得 GRM-RT 模型能够为这些领域的数据分析提供强有力的工具。通过整合数据，模型不仅提升了研究的深度，还拓展了传统测量模型的适用范围，展示了广阔的应用前景。

虽然本研究取得了一些有意义的结果，但是仍然存在一些不足。主要包括三个方面，首先是模型仅对 5 个难度等级和 3 个难度等级计分的情况进行了分析，其他计分等级下模型的性能尚未得到验证。其次，模型目前仅考虑了作答时间，而未加入其他类型的过程性数据进行建模。第三，在考虑作答时间与难度等级之间的相关性时，本研究仅在“中等难度等级”与时间强度参数之间设定了相关关系。然而，在实际测验中，可能存在“难度等级越高，与时间强度参数的相关性越强”的情形；也有可能只有“最高难度”与时间强度参数存在显著相

关，而其他难度等级与时间参数之间的关联较弱。这些情况在本研究中尚未纳入分析。因此，未来可以针对以下四个方面进行深入研究，(1)探索不同先验分布和参数相关性下的模型性能，例如有些学者认为在考生不认真作答的情况下，考生的能力参数和速度参数之间的相关性是轻微的负相关，而不是假设能力和速度之间存在正相关。另一方面，在其它测验情境中，可以使用非信息先验增加模型的适用性，使研究者能够在面对不同形式的数据时，更灵活地调整模型假设。通过对先验分布和参数相关性设定的全面对比与检验，可以为后续研究提供更加稳健、灵活的模型框架。(2)可以考虑引入更多的过程性数据，如考生的鼠标轨迹和鼠标点击次数(De Boeck & Scalise, 2019; Qiao et al., 2023)，眼动数据(Man & Harring, 2019; Man & Harring, 2021)等，上述研究在基于 2PL 模型和作答时间模型的基础上，额外加入了第三个维度的数据，构建了一个联合三个维度的综合模型，从而实现了对作答行为的更为全面的分析。未来的研究可以进一步结合多级计分模型和作答时间模型，并同时纳入更多类型的过程性数据，如鼠标轨迹、点击次数及眼动数据等，形成更加多元和精准的作答行为模型。(3)除了考虑使用原有的对数时间正态模型，不同的作答时间的模型可以应用于未来的研究中，例如可以使用 BOX-COX 正态作答时间模型(Entink et al., 2009)来拟合作答时间数据，BOX-COX 正态作答时间模型允许作答时间数据转换正态分布时具有更大的灵活性。(4) 在现有的 GRM-RT 模型基础上，可以进一步将其扩展为混合模型。通过引入随机效应、群体差异以及多维度 IRT 模型等因素，研究者可以将考生群体按照某些特征（如能力水平、答题策略或其他背景信息）分成不同组别，并分别估计这些组别的参数差异，以期更全面地考察考生的作答模式。多维度 IRT 模型的加入还能在同一模型框架下捕捉到考生多个潜在能力维度间的相互作用，为测评实践提供更具深度与广度的洞察。

参考文献

Beck, M. F., Albano, A. D., & Smith, W. M. (2019). Person-fit as an index of inattentive responding: A comparison of methods using polytomous survey data. *Applied Psychological Measurement, 43*(5), 374－387.

Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*.

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics, 7*(4), 434-455.

Cheng, Q. (2008).*The establishment and application research on three-parameter graded response model*.(Master's thesis), Jiangxi Normal University.

[陈青. (2008).*三参数等级反应模型（3P-GRM）的建立及其应用研究*(硕士学位论文),江西师范大学.]

Cheng, Y., & Shao, C. (2022). Application of change point analysis of response time data to detect test speededness. *Educational and psychological measurement*, *82*(5), 1031-1062.

Choe, E. M., Kern, J. L., & Chang, H. H. (2018). Optimizing the use of response times for item selection in computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, *43*(2), 135-158.

Dai, S., Vo, T. T., Kehinde, O. J., He, H., Xue, Y., Demir, C., & Wang, X. (2021, September). Performance of polytomous IRT models with rating scale data: An investigation over sample size, instrument length, and missing data. In *Frontiers in Education (Vol. 6, p. 721963)*. Frontiers Media SA.

De Boeck, P., & Scalise, K. (2019). Collaborative problem solving: Processing actions, time, and performance. *Frontiers in psychology*, *10*, 1280.

Dunn, T. G., Lushene, R. E., & O'Neil, H. F. (1972). Complete automation of the MMPI and a study of its response latencies. *Journal of Consulting and Clinical Psychology, 39*(3), 381.

Entink, R. K., van Der Linden, W. J., & Fox, J. P. (2009). A Box–Cox normal model for response times. *British Journal of Mathematical and Statistical Psychology*, *62*(3), 621-640.

Fox, J.-P., Entink, K. H. R., & van der Linden, J. W. (2007). Modeling of responses and response times with the package CIRT. *Journal of Statistical Software, 20*(7), 1 – 14.

Fox, J. P., & Marianti, S. (2016). Joint modeling of ability and differential speed using responses and response times. *Multivariate Behavioral Research, 51*(4), 540-553.

Guo, L., Shang, P. L., & Xia, L. X. (2017). Advantages and illustrations of application of response time model in psychological and educational testing. *Advances in Psychological Science, 25*(4), 701-712.

[郭磊，尚鹏丽，夏凌翔. (2017). 心理与教育测验中反应时模型应用的优势与举例. *心理科学进展, 25*(4), 701-712.]

Guo, Y. Y. (2020) *Development and application of polytomously-scored IRT model with response time*(Master's thesis), Jiangxi Normal University.

[郭莹莹. (2020). *融合反应时的多级评分 IRT 模型开发及其应用研究*(硕士学位论文), 江西师范大学.]

Han, S., & Kang, H. A. (2021). Sequential Monitoring of Aberrant Test-Taking Behaviors Based on Response Times. In *Quantitative Psychology: The 85th Annual Meeting of the Psychometric Society, Virtual* (pp. 69-80). Springer International Publishing.

Holden, R. R., & Fekken, G. C. (1993). Can personality test item response latencies have construct validity? Issues of reliability and convergent and discriminant validity. *Personality and Individual Differences, 15*(3), 243-248.

Hung, L. F., & Wang, W. C. (2012). The generalized multilevel facets model for longitudinal data. *Journal of Educational and Behavioral Statistics, 37*(2), 231-255.

Kieftenbeld, V., & Natesan, P. (2012). Recovery of graded response model parameters: A comparison of marginal maximum likelihood and Markov chain Monte Carlo estimation. *Applied Psychological Measurement, 36*(5), 399-419.

LaHuis, D. M., Clark, P., & O'Brien, E. (2011). An examination of item response theory item fit indices for the graded response model. *Organizational research methods*, *14*(1), 10-23.

Liu, Y., & Wang, W. (2024). What Can We Learn from a Semiparametric Factor Analysis of Item Responses and Response Time? An Illustration with the PISA 2015 Data. *psychometrika*, *89*(2), 386-410.

Liu, Z. Y., (2024). *Modeling and application of response time-based mixture item response theory (RTs-mGPCM) in the context of psychology and education* (Master's thesis), Guizhou Normal University.

[刘子瑜. (2024). *在心理与教育视域下结合反应时的混合项目反应理论（RTs-mGPCM）建模及应用* (硕士学位论文), 贵州师范大学.]

Ma, W., & de la Torre, J. (2016). A sequential cognitive diagnosis model for polytomous responses. *British Journal of Mathematical and Statistical Psychology*, *69*(3), 253-275.

Man, K., & Harring, J. R. (2019). Negative binomial models for visual fixation counts on test items. *Educational and Psychological Measurement*, *79*(4), 617-635.

Man, K., & Harring, J. R. (2021). Assessing preknowledge cheating via innovative measures: A multiple-group analysis of jointly modeling item responses, response times, and visual fixation counts. *Educational and Psychological Measurement*, *81*(3), 441-465.

Manapat, P. D., & Edwards, M. C. (2022). Examining the robustness of the graded response and 2-parameter logistic models to violations of construct normality. *Educational and Psychological Measurement, 82*(5), 967-988.

Marianti, S., Fox, J.-P., Avetisyan, M., Veldkamp, B. P., & Tijmstra, J. (2014). Testing for aberrant behavior in response time modeling. *Journal of Educational and Behavioral Statistics, 39*(6), 426–451.

Molenaar, D., Bolsinova, M., & Vermunt, J. K. (2018). A semi-parametric within-subject mixture approach to the analyses of responses and response times. *British Journal of Mathematical and Statistical Psychology*, *71*(2), 205-228.

Molenaar, D., Dolan, C.V. & de Boeck, P. (2012). The heteroscedastic graded response model with a skewed latent trait: Testing statistical and substantive hypotheses related to skewed item category functions. *Psychometrika, 77*(3), 455-478.

Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. (2015). A generalized linear factor model approach to the hierarchical framework for responses and response times. *British Journal of Mathematical and Statistical Psychology*, *68*(2), 197-219.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied psychological measurement*, *16*(2), 159-176.

Naumenko, O. (2014). Comparison of various polytomous item response theory modeling approaches for task-based simulation CPA exam data. AICPA 2014 summer internship project.

Patton, J. M. (2015). *Some consequences of response time model misspecification in educational measurement*. University of Notre Dame.

Qiao, X., Jiao, H., & He, Q. (2023). Multiple-group joint modeling of item responses, response times, and action counts with the conway-maxwell-poisson distribution. *Journal of Educational Measurement*, *60*(2), 255-281.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Society*.

Schnipke, D. L., & Scrams, D. J. (2005). Exploring issues of examinee behavior: Insights gained from response-time analyses. In *Computer-based testing* (pp. 237-266). Routledge.

Shao, C. (2016). *Aberrant response detection using change-point analysis* (Doctoral dissertation). University of Notre Dame.

Sinharay, S., & Johnson, M. S. (2020). The use of item scores and response times to detect examinees who may have benefited from item preknowledge. *British Journal of Mathematical and Statistical Psychology, 73*(3), 397-419.

Sun, J., Xin, T., Zhang, S., & de la Torre, J. (2013). A polytomous extension of the generalized distance discriminating method. *Applied Psychological Measurement*, *37*(7), 503-521.

Tian, Y. S., Zhan, P. D., & Wang, L. J. (2023). Joint cognitive diagnostic modeling for probabilistic attributes incorporating item responses and response times. *Acta Psychilogica Sinica, 55*(9), 1573-1595.

[田亚淑, 詹沛达, 王立君. (2023). 联合作答精度和作答时间的概率态认知诊断模型. *心理学报, 55*(9), 1573-1595.]

Trafimow, D., Wang, T. H., & Wang, C. (2019). From a Sampling Precision Perspective, Skewness Is a Friend and Not an Enemy!. *Educational and Psychological Measurement, 79*(1), 129-150.

Ulitzsch, E., Pohl, S., Khorramdel, L., Kroehne, U., & von Davier, M. (2022). A response-time-based latent response mixture model for identifying and modeling careless and insufficient effort responding in survey data. *Psychometrika*, *87*(2), 593-619.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics, 31*(2), 181-204.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72*(3), 287-308.

van Der Linden, W. J. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics*, *33*(1), 5-20.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing, 27*, 1413-1432.

Veldkamp, B. P., Avetisyan, M., Weissman, A., & Fox, J-P. (2017). Stochastic programming for individualized test assembly with mixture response time models. Computers in Human Behavior, 76, 693-702.

Wang, C., Chang, H. H., & Douglas, J. A. (2013). The linear transformation model with frailties for the analysis of item response times. *British Journal of Mathematical and Statistical Psychology, 66*(1), 144-168.

Wang, C., Weiss, D. J., & Su, S. (2019). Modeling response time and responses in multidimensional health measurement. *Frontiers in psychology*, *10*, 51.

Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology*, *68*(3), 456-477.

Watanabe, S., & Opper, M. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research, 11*(12).

Yu, X., & Cheng, Y. (2019). A change-point analysis procedure based on weighted residuals to detect back random responding. *Psychological Methods*, *24*(5), 658.

Zhan, P. D., Jiao, H., Man, K. W. (2020). The multidimensional log-normal response time model: An exploration of the multidimensionality of latent processing speed. *Acta Psychilogica Sinica, 52*(9), 1132-1142.

[詹沛达, Jiao, H., J., & Man, K. W. (2020). 多维对数正态作答时间模型：对潜在加工速度多维性的探究. *心理学报, 52*(9), 1132-1142.]

Zhong, X. Y., Yu, X. F., Miao, Y., Qin, C. Y., Peng, Y. F., &Tong, H. (2022). Exploration of change point analysis in detecting speededness based on response time data with known/unknown item parameters. *Acta Psychilogica Sinica, 54*(10), 1277-1292.

[钟小缘,喻晓锋,苗莹,秦春影,彭亚风,童昊.(2022).基于作答时间数据的改变点分析在检测加速作答中的探索——已知和未知项目参数.*心理学报, 54*(10), 1277-1292.]

Zhu, H., Jiao, H., Gao, W., & Meng, X. (2023). Bayesian Change-Point Analysis Approach to Detecting Aberrant Test-Taking Behavior Using Response Times. *Journal of Educational and Behavioral Statistics, 48*(4), 490-520.

Development and Application of Graded Response Model Incorporating Response Times

Liu Zhicheng[1], Qin Chunying[1,2], Luo Zhaosheng[1], Yu Xiaofeng[1], Peng Yafeng[1]

(1) School of Psychology, Jiangxi Normal University

(2) School of Mathematics and Information Sciences, Nanchang Normal University

Abstract

With the extensive implementation of computer-based assessments, educators are now able to gather process data in addition to traditional observed responses. Among these process data, response time is of particular significance as it reflects examinees' responding speed, cognitive processes, and degree of engagement with the test items. As information technology keeps advancing, assessment results are no longer confined to conventional item scores. By considering response time, researchers and practitioners can obtain a dynamic and multidimensional view of how examinees behave and interact with the assessment, thus deepening our understanding of test-taking processes. Previous research on Item Response Theory (IRT) models incorporating response time has mainly focused on dichotomous (0 and 1) scoring both domestically and internationally. However, in practical testing situations, such as multiple-choice items, constructed-response tasks, and essays, polytomous (graded) scoring is commonly used. To fill this research gap, the present study expands the Graded Response Model (GRM) by integrating response-time information. Consequently, the GRM - RT model for polytomous data is proposed, and the parameters are estimated using the Hamiltonian Monte Carlo (HMC) method.

To verify the accuracy and robustness of the GRM-RT model, a simulation study was carried out to examine the precision of parameter estimates. The simulation adopted a 2×2×3 factorial experimental design, with the independent variables being examinee ability distribution, sample size, and test length. To evaluate the estimation accuracy, evaluation metrics like Root Mean Square Error (RMSE) and Bias were employed. Regarding the convergence of parameter estimates, it was assessed using trace plots of the Markov chains and the potential scale reduction factor ($\hat{R}$). The results showed that the estimates of examinee parameters were satisfactory, consistently achieving desirable precision levels even under the conditions of small sample sizes and short test lengths. This indicates that the estimation method remains reliable and stable even when resources are

limited. Item parameter estimation also showed high precision, with minimal variation across different conditions under normal distribution, mainly affected by random error. Under skewed distribution conditions, the precision of parameter estimates remained high, suggesting that the model is robust and reliable across various distribution scenarios.

In the empirical study, a mathematics literacy scale in the field of education was analyzed. Model fit was evaluated by means of the Widely Applicable Information Criterion (WAIC) and Leave-One-Out cross-validation (LOO). The results demonstrated that the GRM-RT model had a better model fit than the RTs-mGPCM model. Additionally, the estimates of the scale's item parameters revealed that the model had lower measurement error, further highlighting the excellent psychometric performance of the GRM-RT model and its promising applications in the educational field.

The above findings suggest that the GRM-RT model has great application potential and can serve as a powerful tool for data analysis in various fields of education and psychology.

Keywords Polytomous Scoring, Graded Response Model, Response Times, Modeling

附录一

表 A-1 数学素养调查问卷

	具体描述
项目 1	通过火车时间表计算出从 A 地到 B 地所需时间。
项目 2	计算加税后一台电脑会比卖价贵多少。
项目 3	已知瓷砖和房间面积的情况下计算需要多少块瓷砖才能够铺满某一房间。
项目 4	理解文章中的数学表格。
项目 5	计算一元二次方程，例如$6x^2+5=29$
项目 6	根据 1:10000 比例绘制的地图计算出某两地间的距离。
项目 7	计算出类似 $2(x+3)=(x+3)(x-3)$ 的方程。
项目 8	了解负数的运算，如$(-3)+5=2$。
项目 9	计算一元一次方程，例如$5x+9=15$。
项目 10	判断两个三角形是否全等。
项目 11	掌握勾股定理以判断三角形是否为直角三角形。
项目 12	用枚举法求概率，如一个袋子里装有一个白球和两个红球,随机拿出两个球皆为红球的概率是多少？
项目 13	掌握平行四边形的定义。
项目 14	判断图形是否对称。
项目 15	掌握圆周长和弧长的对应关系，已知圆周长为 $2\pi R$，圆心角为 40°时弧长应为多少?
项目 16	计算一元一次不等式，如$26+2x\geq 30$。
项目 17	掌握直线、线段和射线的区别。
项目 18	计算一元一次方程组,已知$x+y=50$，$3y-x=30$分别求出 x 和 y 的值。
项目 19	掌握平方差公式，即$a^2-b^2=(a+b)(a-b)$。
项目 20	可以对积的乘方进行计算，例如$5^3+5^2=55$

附录二

参数估计使用的是自编的 R 语言程序，具体的 stan 模型代码如下：

```
stan_code <- "
data {
    int<lower=1> N; // number of persons
    int<lower=1> m; // number of items
    int<lower=2> K; // number of categories
    int<lower=1, upper=K> Y[N, m]; // responses
    matrix[N, m] RT; // response times
    real D; // scaling constant
    real rho;
    real sd_speed;
    real intercept; // intercept for Beta
    real z_a; // coefficient for a
    real z_b; // coefficient for b_mid
    real var_e; // variance of error term
}
parameters {
    vector[N] theta; // ability parameters
    vector[N] tau; // speed parameters
    vector<lower=0>[m] a; // discrimination parameters
    ordered[K - 1] b[m]; // item thresholds
    vector<lower=1.5>[m] Alpha; // RT parameters
    vector[m] Beta; // RT parameters
}
transformed parameters {
    real mat_speed = sd_speed * sqrt(1 - rho^2);
    vector[m] b_mid; // middle threshold for each item

    for (j in 1:m) {
        b_mid[j] = b[j, (K - 1) / 2];
    }
}
model {
    // Priors
    theta ~ normal(0, 1);
    tau ~ normal(rho * sd_speed * theta, mat_speed);
    a ~ lognormal(0, 0.5);
    Alpha ~ normal(2, 0.167);

    // Prior for Beta
```

```
    for (j in 1:m) {
        Beta[j] ~ normal(4, 0.3);
    }

    // Prior on thresholds
    for (j in 1:m) {
        b[j, 1] ~ normal(-1.5, 0.5);
        for (k in 2:(K - 1)) {
            (b[j, k] - b[j, k - 1]) ~ normal(1, 0.2);
        }
    }

    // Likelihood
    for (i in 1:N) {
        for (j in 1:m) {
            vector[K] P; // probabilities for each category
            vector[K + 1] P_star; // cumulative probabilities

            // Compute cumulative probabilities P_star
            P_star[1] = 1;
            for (k in 2:K) {
                P_star[k] = inv_logit(D * a[j] * (theta[i] - b[j, k - 1]));
            }
            P_star[K + 1] = 0;

            // Compute probabilities for each category P
            for (k in 1:K) {
                P[k] = P_star[k] - P_star[k + 1];
            }

            // Responses
            target += categorical_lpmf(Y[i, j] | P);

            // RTs
            real mu_ij = Beta[j] - Alpha[j] * tau[i];
            real sigma_j = 1 / Alpha[j];
            RT[i, j] ~ lognormal(mu_ij, sigma_j);
        }
    }
}
"
```